MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

A.I. Memo No. 1333 December, 1991

On the Uniqueness of Correspondence under Orthographic and Perspective Projections

Ronen Basri

Abstract

The task of shape recovery from a motion sequence requires the establishment of correspondence between image points. The two processes, the matching process and the shape recovery one, are traditionally viewed as independent. Yet, information obtained during the process of shape recovery can be used to guide the matching process. This paper discusses the mutual relationship between the two processes. The paper is divided into two parts. In the first part we review the constraints imposed on the correspondence by rigid transformations and extend them to objects that undergo general affine (non rigid) transformation (including stretch and shear), as well as to rigid objects with smooth surfaces. In all these cases corresponding points lie along epipolar lines, and these lines can be recovered from a small set of corresponding points. In the second part of the paper we discuss the potential use of epipolar lines in the matching process. We present an algorithm that recovers the correspondence from three contour images. The algorithm was implemented and used to construct object models for recognition. In addition we discuss how epipolar lines can be used to solve the aperture problem.

This report describes research done at the Massachusetts Institute of Technology within the Artificial Intelligence Laboratory. Support for the laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-85-K-0124. Ronen Basri is supported by the McDonnell-Pew and the Rothchild postdoctoral fellowships.

1 Introduction

Correspondence is a process of relating information in one image to its equivalent in others. Using correspondence a vision system can infer the 3-D structure of the observed scene, an inference that is significantly more difficult to make from a single 2-D image. The establishment of correspondence is itself a difficult task. Various methods were developed in recent years to achieve correspondence for stereo vision (e.g., [Marr and Poggio 1979, Grimson 1980, Baker and Binford 1981]), motion analysis (e.g., [Koenderink and Van Doorn 1975, Ullman 1979, Longuet-Higgins 1981, Hildreth 1984]), and object recognition (e.g., [Fischler and Bolles 1981, Grimson and Lozano-Pérez 1984, Lowe 1987, Huttenlocher and Ullman 1987]).

In motion analysis the stage of establishing correspondence is usually viewed as independent of the stage of shape recovery [Ullman 1978]. According to this view, the correspondence is determined so as to minimize the observed 2-D motion along the image sequence. No assumptions are made at this stage with respect to the shape of the moving objects or to the transformations they undergo. In this way correspondence can be found even for objects that undergo non rigid transformations, and when the images contain a number of objects moving differently.

The distinction between the two processes of correspondence and shape recovery is useful when the motion between the frames is relatively small, in which case a minimization process can resolve the correspondence correctly. When, however, "long range motion" is considered, minimization techniques often fail to find the correct correspondence. Information about the transformation may be used in these cases to guide the process of establishing correspondence.

An important application that requires correspondence under "long range motion" conditions is the construction of 3-D representations for object recognition. In this process shape information is accumulated over time until a complete model is constructed for the object. During this period the object may be observed in positions that significantly differ from one another. Yet, it is desired for this process to tolerate such differences.

Correspondence is not only useful for constructing object-centered models, but also for viewer-centered ones. Recognition schemes that use viewer-centered representations were recently developed. In [Ullman and Basri 1991] an object is represented by a small number of its 2-D images together with the correspondence between the images. The appearance of an object from different viewpoints is predicted by the linear combinations of its model images. These predictions are exact for rigid objects. Similar representations were used by Poggio and Edelman [1990]. Their approach approximates the appearance of objects from arbitrary viewpoints using radial basis functions.

Point-to-point correspondence between images is therefore crucial for constructing

both object-centered as well as viewer-centered representations. In the object-centered case this follows from the fact that structure from motion algorithms require full correspondence between the images. Once the correspondence is known structure recovery is fairly straightforward. In the viewer-centered case full correspondence between images provides implicit information about the depth values of the points. The stability of a representation, measured by the errors induced when the appearance of the modeled object is predicted from arbitrary viewpoints, tends to increase as the images used to construct the models are taken from viewing angles that are relatively distant from one another.

One assumption that is generally used in different vision applications such as motion and object recognition is that the objects observed are rigid. Lee and Huang [1990] have recently addressed the question of how rigidity affects the solution to the correspondence problem. They showed that under an orthographic projection the correspondence to points can only be determined up to straight lines (known as "the *epipolar lines* of the points"), and that four corresponding points determine the position of these lines. They did not specify any method to resolve the correspondence within the lines.

The epipolar line idea is not new. It is extensively used in stereopsis, but rarely used in establishing correspondence in motion analysis. Bolles and Baker [1985] used epipolar lines to analyze motion sequences obtained by a translation along a straight line. Yachida [1986] and Ayache and Lustman [1987] used it in developing their trinocular stereovision algorithm. In this paper we examine the use of epipolar lines in establishing correspondence for depth reconstruction. In the first part of this paper (Section 2) we review the theory behind epipolar lines and how to compute them from a small number of corresponding points. The formulation we use is somewhat different from that presented by Lee and Huang [1990], and we analyze the similarity and the differences between orthographic and perspective projection models. We show that epipolar lines exist even in more complicated situations, such as when an object undergoes a general linear transformation (including stretch and shear), and when objects with smooth bounding surfaces are considered. In the second part of this paper (Section 3) we show that the correspondence is not determined uniquely even when three or more images are given. Additional images can, however, be used to heuristically resolve the correspondence [Yachida 1986]. We have applied this method to arbitrarily curved images, and used the results to construct object models for recognition. In addition we discuss how epipolar lines can be used to solve the aperture problem.

2 Correspondence from Two Images

The correspondence problem discussed below is defined as follows. Given a pair of 2-D images, for every point in space that is projected to both images find its location in the two

images. Often only feature points (such as contour points) are considered. We examine this problem assuming the images differ by a rigid transformation. We consider two projection models, orthographic projection (with a uniform scale factor to compensate for depth changes) and perspective projection. We begin our discussion by introducing general properties for both projection models, and later prove these properties for each of the models separately. Finally, we extend these properties to more complicated cases, such as objects that undergo general affine transformations (rather than rigid ones) and objects with smooth bounding surfaces.

Our analysis consists of three steps:

1. We show that rigidity divides the images into sets of epipolar lines. Their correspondence is determined by the transformation that separates the images, but the correspondences of points along the lines cannot be determined.

2. The epipolar lines can be recovered from a small set of corresponding points, four in the orthographic case and seven in the perspective case.

3. These results apply also to objects that undergo general affine transformation and to objects with smooth bounding surfaces.

Proposition 1 establishes that in a pair of images related by a rigid transformation a point in one image can potentially match in the second image any point that lies along a straight line (which is referred to as "the *epipolar line* of that point").

Let P_1 and P_2 be two projections (either orthographic or perspective) of a rigid object from two given viewpoints. Let (x, y) be the projection of some object point in P_1.

Proposition 1: The corresponding point to (x, y) in P_2 lies along a straight line given by:

$$(x', y') = \mathrm{u} + \alpha(z)\mathrm{v}$$

where $\mathrm{u}, \mathrm{v} \in \mathcal{R}^2$ are constants (namely, independent of z), and α is a scalar function of z.

Following Proposition 1, given the transformation that relates the two images, the correspondence is determined up to a straight line. The vectors u and v are determined both by the transformation and by the 2-D position of the point (x, y), while α is the only component that depends on z, the depth value of the point in 3-D. There is a one-to-one mapping between the position of p along the epipolar line in P_2 and its depth value. Every different depth value corresponds to a different location of p along the epipolar line, and every different location along the epipolar line determines a different depth value.

In some cases epipolar lines vanish and point correspondence is uniquely determined. This occurs in the degenerated case when v = 0. In this case the position (x', y') does not depend on the depth value of the point. Under orthographic projection this occurs when the object is rotated around the line of sight and then translated arbitrarily. Under perspective projection v vanishes when the object is rotated around the camera.

The epipolar lines are parallel in the orthographic case, since in this case v depends solely on the transformation and is therefore common to all object points. This is not always true in the perspective case. In this case the epipolar lines are parallel only if the transformation includes no translation in depth. If, however, $t_z \neq 0$ the epipolar lines coincide at a single point known as the *focus of expansion.*

A rigid transformation divides the image into epipolar lines within which correspondence cannot be determined. Every epipolar line in one image has its corresponding epipolar line in the second image, in the sense that, all the points that lie along some epipolar line in the first image share the same epipolar line in the second image and vice versa. This is established in Proposition 2.

Proposition 2: Let $p_1, p_2 \in P_1$ be two points that lie along some common epipolar line. The epipolar line of p_1 and the epipolar line of p_2 in P_2 coincide.

Since rigidity alone does not determine the correspondence except up to epipolar lines, it may in some cases be sufficient to recover the epipolar lines rather than all the parameters of the transformation that separates the two images. Interestingly, under orthographic projection the epipolar lines are determined from two images while the transformation is not. Four non coplanar points are required for this task. The transformation breaks up into its planar parts and non planar parts. The planar parts of the transformation are determined by the epipolar lines, while the non planar parts, the rotation in depth, cannot be recovered. In the perspective case both the epipolar lines and the transformation are determined from two images. In this case seven points are required.

The results above apply in two additional cases that extend beyond the set of rigid transformations. Epipolar lines exist when the objects considered undergo general 3-D affine transformation, which includes stretch and shear. The same applies under orthographic projection to objects with smooth bounding surfaces. In this case the contours change their position on the object with the viewpoint. (See a discussion in [Basri and Ullman 1988].) This motion is projected along epipolar lines (See section 2.3 below). In both cases, corresponding points lie along epipolar lines, and these epipolar lines can be recovered from a small set of corresponding points.

2.1 Orthographic Projection

In this section we repeat the results presented in the beginning of Section 2 and prove them for the orthographic case. Let P_1 and P_2 be two images of a rigid object from two arbitrary viewpoints. Let $\mathrm{p} = (x, y, z)$ be an object point, its position in P_1 is given by (x, y), and its position in P_2 is given by (x', y') which is the orthographic projection of $sR\mathrm{p} + \mathrm{t}$, where s is a scale factor, R is a 3×3 rotation matrix, and t is a translation vector.

In the following analysis we assume that the transformation between the images (namely, s, R, and t) is known. We select a point (x, y) in the first image and compute its possible positions in the second image. We show that the set of these positions forms a straight line, and that the exact position along this line is determined by its depth value.

Proposition 1a: Given a rigid transformation defined by $\{s, R, \mathrm{t}\}$ and a point $(x, y) \in P_1$, its corresponding point in P_2 lies along the epipolar line given by:

$$(x', y') = \mathrm{u} + z\mathrm{v}$$

where $\mathrm{u}, \mathrm{v} \in \mathcal{R}^2$ are constants.

Proof: Denote r_{ij}, $(1 \leq i, j \leq 3)$ the elements of R, t_x and t_y the horizontal and vertical components of the translation, since

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} s(r_{11}x + r_{12}y + r_{13}z) + t_x \\ s(r_{21}x + r_{22}y + r_{23}z) + t_y \end{pmatrix}$$

we define

$$\begin{aligned} \mathrm{u} &= \begin{pmatrix} sr_{11}x + sr_{12}y + t_x \\ sr_{21}x + sr_{22}y + t_y \end{pmatrix} \\ \mathrm{v} &= \begin{pmatrix} sr_{13} \\ sr_{23} \end{pmatrix} . \end{aligned}$$

Notice that since the transformation is given, u and v are determined for a particular point (x, y), and consequently its corresponding point lies along the straight line $\mathrm{u} + z\mathrm{v}$. When the depth value, z, of the point is given, the location of the corresponding point along the line is determined, and vice versa, selecting a corresponding point along the line determines its depth value. When $\mathrm{v} = 0$ the epipolar line vanishes into a point. In this case the images are separated by a rotation about the line of sight (plus some arbitrary translation). For symmetry reasons we obtain the same results for points in the second image, namely, that their corresponding points in P_1 lie along straight lines.

The epipolar lines in each of the images are parallel. This follows from the fact that v depends solely on the transformation, and therefore has a common value for all image points. All the points in P_1 that lie along a single epipolar line share the same epipolar line in P_2. This is established in the following Proposition.

Proposition 2a: Let $p_1, p_2 \in P_1$ be two points that lie along some common epipolar line. The epipolar line of p_1 and the epipolar line of p_2 in P_2 coincide.

Proof: All the epipolar lines are parallel. According to the definition of an epipolar line, since p_1 and p_2 lie along a single epipolar line, both are possible matches of a single point, q, in P_2. Therefore, the epipolar lines of p_1 and p_2 intersect in q, and since epipolar lines are parallel they must coincide. Consequently, rigidity determines the correspondence between epipolar lines, but does not resolve the correspondence within these lines.

When only two images are given the transformation cannot be fully recovered. The epipolar lines, however, can be recovered using a correspondence set of four non coplanar points. A linear equation from which the epipolar lines can be computed is given below. We shall use the following notation. Let $(x_i, y_i) \in P_1$ and $(x'_i, y'_i) \in P_2$ be a pair of corresponding points, namely, they are the projections of a common point in 3-D space, $p_i = (x_i, y_i, z_i)$. We shall have n such correspondences. (To solve this equation n must be ≥ 4.) Denote $\mathbf{x} = (x_1, ..., x_n)$, $\mathbf{y} = (y_1, ..., y_n)$, $\mathbf{z} = (z_1, ..., z_n)$, $\mathbf{x}' = (x'_1, ..., x'_n)$, $\mathbf{y}' = (y'_1, ..., y'_n)$, and $\mathbf{1} = (1, ..., 1) \in \mathcal{R}^n$. According to [Ullman and Basri 1991], $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$, $\mathbf{x}'$, $\mathbf{y}'$, and $\mathbf{1}$ are all embedded in a 4-D linear space. This follows from the identities below

$$\begin{aligned} \mathbf{x}' &= sr_{11}\mathbf{x} + sr_{12}\mathbf{y} + sr_{13}\mathbf{z} + t_x\mathbf{1} \\ \mathbf{y}' &= sr_{21}\mathbf{x} + sr_{22}\mathbf{y} + sr_{23}\mathbf{z} + t_y\mathbf{1} \end{aligned}$$

Consequently, $\{\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{1}\}$ span a 4-D linear space to which $\mathbf{x}'$ and $\mathbf{y}'$ also belong. Therefore, there exist nonzero scalars a_1, a_2, b_1, b_2, and c such that:

$$a_1\mathbf{x} + a_2\mathbf{y} + b_1\mathbf{x}' + b_2\mathbf{y}' + c\mathbf{1} = 0$$

These coefficients are determined (up to a scale factor) by four non coplanar points. The epipolar line are immediately derived from this equation. (This result is proved somewhat differently in [Huang and Lee 1989, Lee and Huang 1990].)

The epipolar lines break the transformation that relates the images into its planar components and its non planar ones. The planar components can be recovered from the epipolar lines, while the non planar ones cannot be determined from two images. The translation component perpendicular to the epipolar line is given by c. (The translation

components can be discarded altogether if we consider differences between points rather than the points themselves.) The values of the other coefficients are given below.

$$\begin{aligned} a_1 &= s r_{32} \\ a_2 &= -s r_{31} \\ b_1 &= r_{23} \\ b_2 &= -r_{13} \end{aligned}$$

The scale factor is therefore given by the ratio

$$s = \sqrt{\frac{a_1^2 + a_2^2}{b_1^2 + b_2^2}}$$

The relative angle between the epipolar lines determines the planar parts of the rotation, as explained below. A 3-D rotation can be decomposed into a sequence of three successive rotations: a rotation about the Z-axis by an angle α, a second rotation about the Y-axis by an angle β, and a third rotation about the Z-axis by an angle γ. Under this decomposition the following identities hold

$$\begin{aligned} r_{32} &= \sin\alpha \sin\beta \\ r_{31} &= -\cos\alpha \sin\beta \\ r_{23} &= \sin\beta \sin\gamma \\ r_{13} &= \sin\beta \cos\gamma \quad . \end{aligned}$$

We therefore obtain that

$$\begin{aligned} \alpha &= \tan^{-1}\frac{a_1}{a_2} \\ \gamma &= -\tan^{-1} -\frac{b_1}{b_2} \end{aligned}$$

while β cannot be determined.

We can visualize this decomposition in the following way. After compensating for the translation and scale changes, we first rotate the image P_1 by α. Consequently, the epipolar lines point in P_1 to a horizontal direction. We then rotate the second image, P_2, by $-\gamma$. As a result, the epipolar lines in P_2 also point horizontally. The images obtained are related by a rotation about the vertical axis, which is a rotation in depth. Following such a rotation the points move horizontally, which is, along the (rotated) epipolar lines. This motion cannot be recovered since it depends both on the angle of rotation, β, and on the depth of the points.

An essentially similar break up of the transformation was suggested by Ullman [1983]. In his proof, however, a correspondence set of five points was required to recover the planar parts of the transformation. We can see here that four non coplanar points are sufficient, since the epipolar lines can be recovered from four such points, and the break up is completely described by the epipolar lines.

2.2 Perspective Projection

In this section we repeat the results presented in the beginning of Section 2 and prove them for the perspective case. We use the following notation. An object point p is denoted by (zx, zy, z). It is projected in P_1 to the position (x, y) and in P_2 to (x', y'). (There the actual 3-D position of the point is denoted by $(z'x', z'y', z')$.)

Proposition 1b: Given a rigid transformation that includes a rotation R and a translation t, and given a point $(x, y) \in P_1$, its corresponding point in P_2 lies along the epipolar line given by

$$(x', y') = \mathrm{u} + \alpha(z)\mathrm{v}$$

where $\mathrm{u}, \mathrm{v} \in \mathcal{R}^2$ are constants, and α is a scalar function of z.

Proof: Denote

$$\begin{pmatrix} x_r \\ y_r \\ z_r \end{pmatrix} = R \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}$$

Note that x_r, y_r, and z_r are all independent of z. Now, since

$$z' \begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = Rz \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} + \mathrm{t} = z \begin{pmatrix} x_r \\ y_r \\ z_r \end{pmatrix} + \mathrm{t}$$

we obtain that

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \frac{1}{z_r} \begin{pmatrix} x_r \\ y_r \end{pmatrix} + \frac{1}{z'} \begin{pmatrix} t_x - t_z \frac{x_r}{z_r} \\ t_y - t_z \frac{y_r}{z_r} \end{pmatrix}$$

and so we define

$$\begin{aligned} \mathrm{u} &= \frac{1}{z_r} \begin{pmatrix} x_r \\ y_r \end{pmatrix} \\ \mathrm{v} &= \begin{pmatrix} t_x - t_z \frac{x_r}{z_r} \\ t_y - t_z \frac{y_r}{z_r} \end{pmatrix} \\ \alpha(z) &= \frac{1}{z'} = \frac{1}{z z_r + t_z} \quad . \end{aligned}$$

Parallel epipolar lines are obtained when $t_z = 0$. In this case v is independent of the position of the point and depends solely on the transformation. If, however, $t_z \neq 0$ the epipolar lines intersect in one point, called the *focus of expansion*. This point stands for $z = 0$, and its location in P_2 is given by

$$\begin{pmatrix} x'_0 \\ y'_0 \end{pmatrix} = \frac{1}{t_z} \begin{pmatrix} t_x \\ t_z \end{pmatrix}$$

The location of the focus of expansion in P_1 corresponds to the case when v = 0. This condition implies the following linear equation system

$$\begin{aligned} t_z x_r &= t_x z_r \\ t_z y_r &= t_y z_r \end{aligned}$$

from which this location can be retrieved. (Recall that x_r, y_r, and z_r are linear functions of x and y.)

Similar to the orthographic case, points that lie on a common epipolar line in one image share the same epipolar line in the other.

Proposition 2b: Let $p_1, p_2 \in P_1$ be two points that lie along some common epipolar line. Assume both p_1 and p_2 are not the focus of expansion. The epipolar line of p_1 and the epipolar line of p_2 in P_2 coincide.

Proof: If $t_z = 0$ the epipolar lines are parallel and the proof is identical to that of the orthographic case. If $t_z \neq 0$ the epipolar lines in each image intersect in the focus of expansion. Since the points lie along a common epipolar line in P_1 there exists a point q in P_2 that is a possible match to both points, q is not the focus of expansion. Therefore, the epipolar line of p_1 and that of p_2 intersect in q, and since both lines also intersect in the focus of expansion they must coincide.

In the perspective case the transformation can be determined in general (up to a scale factor) by a correspondence set of seven points [Longuet-Higgins 1981, Tsai and Huang 1984]. There is still no proof for whether this is the minimal number. Roach and Aggarwal [1979] showed by counting the number of unknowns that five points may be sufficient. For the sake of completeness we review in Appendix A one method to recover the transformation from eight corresponding points using essentially linear operations. This method appeared in Tsai and Huang [1984].

It is worth noting that although in the perspective case the transformation can be recovered from two images the computation may in many cases be unstable. This happens when the object is relatively distant from the camera, in which case depth differences are relatively small and perspective distortions are negligible, and when the depth translation components are small, in which case the epipolar lines are nearly parallel. These cases are

essentially similar to the orthographic case. In both cases the transformation obtained is unstable, and a third image may be required to recover the transformation reliably. The epipolar lines, however, remain stable since they depend mainly on those components of the transformation that can be measured reliably.

2.3 Extensions

In the previous discussion we showed that rigidity determines the correspondence up to epipolar lines and that the position of points along these lines is determined by their depth values. We also showed that the epipolar lines can be recovered from a small set of corresponding points. In this section we consider two additional cases to which epipolar lines apply. These cases include images of objects that undergo general affine transformation and contour images of rigid objects with smooth bounding surfaces.

An affine transformation in 3-D space is composed of a general linear transformation followed by a translation. The set of affine transformations contains, in addition to all the rigid transformations, also stretch and shear. Similar to the rigid case, in a pair of images of an object that undergoes an affine transformation, corresponding points lie along epipolar lines. This is true both when the images are orthographic as well as perspective projections of the object. This follows from the fact that in proving the results above we never used the special properties of the rotation matrix.

When a pair of images is given, whether the objects in these images are moving rigidly or whether they undergo an affine (non rigid) transformations is indistinguishable. Basri and Ullman [1991] (see also [Poggio 1990]) showed that under orthographic projection the set of images of a rigid object is contained in a 4-D linear space, and that additional (quadratic) constraints distinguish between these images and other vectors in this space. These other vectors are, in fact, images obtained by applying a general 3-D affine transformation to the object. The quadratic constraints cannot be recovered from two images. Hence, it is impossible to distinguish between the two cases when only two images are given. A similar ambiguity holds under perspective projection. It is worth noting that general affine transformations approximate the way moving objects are observed in movies from different viewpoints. This effect is known since 1859 as the La Gournerie Paradox and was recently discussed by Jacobs [1991].

A second interesting case is that of rigid objects with smooth surfaces. The bounding contours of such an object are generated by surface patches that are tangent to the line of sight. These patches are usually referred to as the *rim* [Koenderink and Van Doorn 1979] or the *contour generator* [Marr 1977] of the object. Since the surface of the object is smooth, when the object rotates in depth a new set of surface patches that are now tangent to the new line of sight replaces the original rim, generating a new set

of bounding contours. Establishing correspondence between the original and the new bounding contours of the object is therefore problematic, since the contours undergo in addition to the rigid transformation also some arbitrary motion that depends on the exact shape of the object.

Tracing the positions of these contours is useful for any shape reconstruction and object recognition scheme that is based on contour matching. A method to predict the appearance of objects with smooth bounding surfaces for recognition was recently developed [Basri and Ullman 1988]. The method assumes an orthographic projection and uses the 3-D curvature of points along the contours to follow their change in position with viewpoint. The curvature values were computed from a few images of the object by matching the contours in these images.

The next observation demonstrates that epipolar lines are useful in determining correspondences between orthographic images of objects with smooth bounding surfaces. We first look at the simpler case of an object that rotates about the vertical axis. Let p be a rim point, and let us take a horizontal section of the object through p. (Namely, if p $= (x_0, y_0, z_0)$ we consider the plane $y = y_0$.) The intersection of the surface of the object with this plane forms a space curve, C. When the object rotates, the rim point p changes its position on the object along C. Denote the new rim point by p$'$. Since this is a rotation about the Y-axis, the epipolar lines in both images are horizontal. Therefore, all the points on C including p and p$'$ are projected to a common epipolar line in both images.

We now extend this observation to general rigid transformations. Rotation is the only component that affects the rim. Translation and scaling do not change the rim and therefore can be disregarded. A 3-D rotation can be decomposed into three successive rotations, around the Z-, Y-, and Z-axes. (The same decomposition used in Section 2.1.) As we did in Section 2.1, we apply the first rotation to the first image, and (the inverse of) the last rotation to the second image. Both rotations are image rotations, and they do not change the rim. They rotate the epipolar lines in the two images into a horizontal direction. (See Section 2.1.) Therefore, after applying these rotations we obtain the situation described above for the simpler case, namely, the two images are related by a rotation about the vertical axis, and their epipolar lines are horizontal. Therefore, the observed position of the rim points change along epipolar lines.

Figure 1 shows the epipolar lines in two orthographic projections of a VW car. Notice that the matching between silhouette points along epipolar lines is good although they are generated by smooth surfaces.

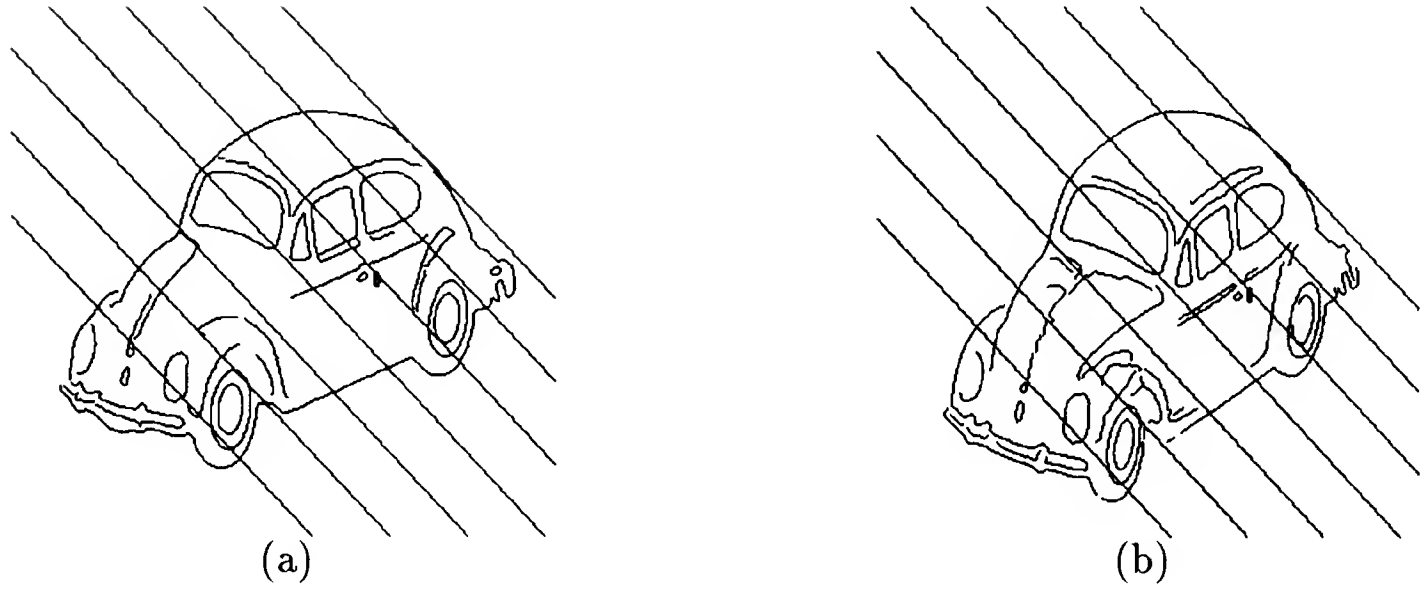


Figure 1: Epipolar lines in two orthographic projections of a VW car. Note the fact that corresponding points lie along epipolar lines. The silhouette contours deserve special attention for being generated from smooth surfaces.

3 Resolving Point Correspondence

In the previous section we have shown that rigidity alone is insufficient to solve the correspondence problem uniquely from two images. It divides the images into epipolar lines, their matching is determined by the transformation that separates the images, but the correspondence of points within the lines cannot be resolved. In this section we examine the problem of establishing correspondence in three or more images. We show that, similar to the case of two images, the correspondence is not determined uniquely. Additional images, however, provide constraints that can be used to solve the problem heuristically (e.g., the trinocular stereovision algorithm [Yachida 1986]). We discuss several additional constraints that can be used together with epipolar lines to find the correspondence between images. These methods were implemented and the results are presented below.

It should be noted that the use of epipolar lines to determine correspondence is limited to those regions in the images that are consistent with a rigid (or affine) transformation. When the images contain a number of rigid objects moving independently each of the objects may determine a different set of epipolar lines. A segmentation process must be applied to separate these objects and divide the images into regions with consistent sets of epipolar lines. We shall not address the segmentation problem in this paper.

3.1 Correspondence from Three Images

We have so far explored the establishment of correspondence from two images. We showed that the correspondence between points in the images cannot be uniquely resolved even if the transformation is known. We now address the following question. Can the correspondence be resolved when three or more images are considered? Structure from motion theory demonstrates that the answer to this question is not trivial. When correspondence is given, under orthographic projection two images are not sufficient to recover the transformation, but three are [Ullman 1979, Huang and Lee 1989]. The correspondence problem is nevertheless different from the structure from motion problem. Point correspondence cannot be resolved by using any number of additional images. Yet, additional images provide information that can be used to filter out less likely solutions.

Proposition 3 establishes that point correspondence cannot be resolved from any number of images. Let $P_1, P_2, ..., P_k$ be k images. Let (x_i, y_i), $1 \leq i \leq k$ be the locations of a point p $= (x, y, z)$ in P_i. (Assume w.l.g. that $x_1 = x$ and $y_1 = y$.) Let T_i, $2 \leq i \leq k$ be the rigid transformation applied to p in P_i, assuming orthographic projection.

Proposition 3: Given $T_2, ..., T_k$, the set of possible locations of p in $P_2, ..., P_k$ forms a straight line in $\mathcal{R}^{(k-1)\times 2}$

$$(x_2, y_2, ..., x_k, y_k) = \mathbf{u} + z\mathbf{v}$$

where $\mathbf{u}, \mathbf{v} \in \mathcal{R}^{(k-1)\times 2}$ are constants.

Proof: This is obtained simply by defining $\mathbf{u} = (u_2, ..., u_k)$ and $\mathbf{v} = (v_2, ..., v_k)$, where $u_i, v_i \in \mathcal{R}^2$ are the corresponding vectors u and v from Proposition 1a.

This proposition implies that the number of possible correspondences for each point is infinite. Every possible assignment of z yields to a different location of the points in all of the images. An equivalent claim can be made in case of perspective projection.

There is, however, one additional consequence to this proposition. Determining the correspondence between two of the images immediately implies the correspondence in all other images. This property suggests a hypothesis-verification heuristic to recover correspondence. The algorithm first selects a point in the first image, hypothesizes its correspondence in the second image, computes accordingly its position in the third, and then verifies its appearance in the predicted location. This algorithm is used in Trinocular stereopsis [Yachida 1986]. The algorithm can be defined in two versions. The first requires the transformation between the images. It predicts the position of points in the third image by explicitly computing their depth values. The second requires the epipolar lines between all pairs of images. It predicts the position of points in the third image by intersecting epipolar lines.

Version 1.

1. Select a point p $= (x, y) \in P_1$ and find its epipolar lines A in P_2.
2. For all candidates $q_1, ..., q_n$ along A compute the corresponding depth value $z_1, ..., z_n$.
3. For every possible depth value, $z_1, ..., z_n$, compute the position of the point (x, y, z_i) in P_3 and verify its actual appearance at this location.

Version 2.

1. Select a point p $= (x, y) \in P_1$, and find its epipolar lines A in P_2 and B in P_3.
2. For all candidates $q_1, ..., q_n$ along A compute their epipolar lines $C_1, ..., C_n$ in P_3.
3. Intersect each of the lines, $C_1, ..., C_n$, with B and verify the actual appearance of p in these locations.

The two versions of the algorithm are essentially similar. The first version uses the transformation between the images to compute depth values. The second version replaces this computation by intersecting epipolar lines. Note that the transformation can be computed from three images using four non coplanar points [Ullman 1979]. The second version can be used only if the epipolar lines C_i intersect with B. The meaning of this requirement is for every image its epipolar lines with respect to the other images should all be non parallel. (So that, if we take for example P_3, its epipolar line with respect to P_1 is not parallel to its epipolar line with respect to P_2, and so forth.) This requirement is equivalent to requiring the transformations to be independent. Unless this condition is met structure-from-motion algorithms cannot recover the transformation from correspondence [Huang and Lee 1989].

One observation following this algorithm is that, since epipolar lines are defined for pairs of images, one can use different sets of anchor points to recover the epipolar lines in each of the pairs. This is different from most existing structure from motion algorithms, which require from the set of anchor points to be identical in all three images.

Note that the use of three images rather than two is reasonable since three images are required to recover structure from motion under orthographic projection [Ullman 1979, Huang and Lee 1989] and to form a viewer-centered representation for a rigid object [Ullman and Basri 1991].

The algorithm handles both rigid objects as well as objects that undergo general 3-D affine transformations. There is, however, some difference between the two cases. When four or more images are considered certain configurations of epipolar lines may be consistent with some affine transformations but with no rigid ones. This is concluded from [Basri and Ullman 1991], since three images are necessary to determine the functional

constraints that distinguish rigid transformations from affine ones. These constraints then restrict the possible configuration of the epipolar lines in larger sets of images.

Stability problems are anticipated in applying the above algorithm when contour pieces are tangential to the epipolar lines. The image in which such an event occurs should be used in this case as the third image. Notice that the three images are symmetric, in the sense that, the algorithm can treat them in any order.

It should be stressed that both versions do not guarantee uniqueness. Occasionally candidates may be found consistent with all three images. Further pruning between these candidates is required. In general the algorithm gives better results for sparse images than for dense ones and for images with arbitrarily distributed texture than for images with uniform texture. (Density refers here to the number of points actually considered by the algorithm relative to the total area of the images.) A common way to reduce the density of an image is to consider its edge map. Edge images are in general still too dense, and a naive implementation of the algorithm would fail to provide a unique solution for many of the points. To avoid this problem we suggest to apply this matching procedure to edges rather than to points, using the assumption that continuous edges tend to remain continuous in all images. Unlike Ayache and Lustman [1987], our implementation is not confined to straight line segments, but is applied to arbitrarily curved ones. We exploit the shape variance of image contours to discriminate between correct and false matches.

The modified algorithm was implemented and run on real images. An example is given in Figures 2-4. In these figures correspondence was sought between three edge images of a VW car (Figure 2). We first selected a contour from the first image. Then we found all the contours in the second image that could possibly match the selected contour. For each of the candidates we computed their location in the third image. We repeated this process for a number of contours. Figure 3 shows the best candidates projected to the third image. Figure 4 shows some of the other candidates projected. None of these candidates match an actual contour (although some of their points do). The results of this algorithm were used to create object models for recognition. An example for the use of these models can be found in [Ullman and Basri 1991].

3.2 Alternative constraints

In this section we briefly discuss several constraints that, combined with the epipolar lines, can be used for establishing point correspondence. The first constraint is traditionally referred to as the *ordering* constraint. Most objects are opaque. Contour segments (and points) on such objects retain their spatial order from different viewpoints. Therefore, a contour segment B that lies between two contour segments, A and C, in one image would in general match some contour segment B′, which lies between the two corresponding

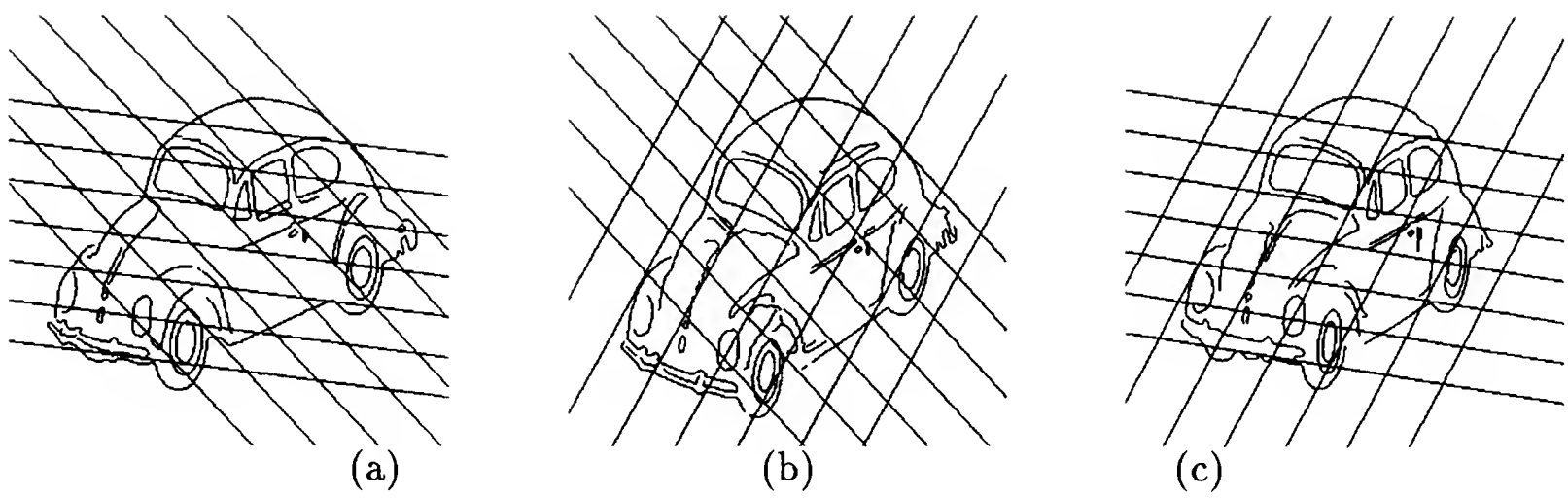
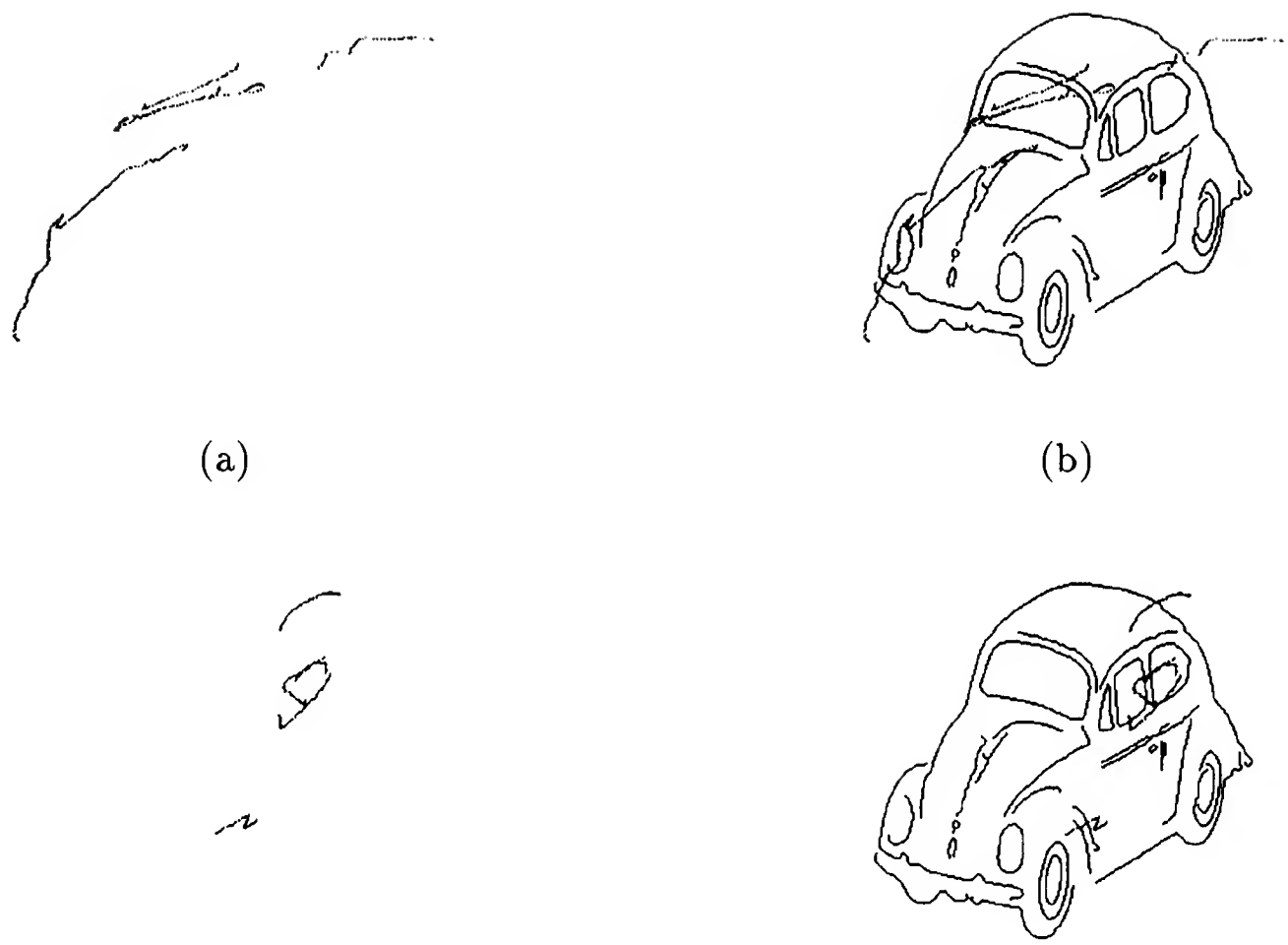


Figure 2: Epipolar lines in three images of a VW car. Every image contains one set of epipolar lines against each of the other two images.

Figure 3: Application of the three images algorithm to four contour pieces selected from the car in Figure 2(a). (The selected contours include the roof silhouette, the front window, the rear side window, and the bottom silhouette.) (a) The best prediction found by the algorithm for the four contour pieces. (b) This prediction overlapped with the actual (third) image.



(a) (b)

(c) (d)

Figure 4: Correspondence candidates that were not selected by the algorithm because their predictions poorly matched the third image. (a) Prediction of false candidates. (b) This prediction overlapped with the actual image. (c) Another prediction of false candidates. (d) This prediction overlapped with the actual image.

contour segments, A′ and C′ respectively. (Notice that right, left, up, and down can still change, as in the case of a 180° rotation around the line of sight.)

Other cues that may be helpful to resolve the correspondence are parallelism and symmetry. If a pair of contour segments are parallel or symmetrical in one image their corresponding contour segments in the second image are often parallel or symmetrical respectively. Resolving the correspondence for one segment would therefore indicate a solution for the other segment. It is worth mentioning, however, that perspective projection does not maintain parallelism, and that symmetrical components often appear skewed in the image under both projections. Incorporating these cues into a process of resolving the correspondence may therefore be fairly difficult.

Epipolar lines can be used to improve the correspondence achieved under aperture conditions. Under these terms matching between contours is given along a direction perpendicular to the contours [Marr and Ullman 1981]. Common techniques to correct the matching use iterative computation to maximize the smoothness of the flow [Hildreth 1984], use sequences of images to find a rigidly consistent solution [Ullman 1984], or compute a smooth, locally affine solution [Burt *et al* 1990, Bachelder and Ullman 1991]. The epipolar line technique offers an exact solution to the aperture problem for full rigid motion that is both computationally simple and resolves the correspondence for as few as two images.

Figure 5 compares the matching obtained under aperture conditions with the matching obtained using epipolar lines for two car silhouettes. It should be noted that in general the aperture problem is associated with short range motion applications. In this case the computation of epipolar lines tends to be unstable. One way to overcome this problem is to recover the epipolar lines for a sequence of images, such that the difference between each pair of consecutive images is small, but the overall transformation accumulated along the sequence is large. Alternatively, if two "distant" images are provided the images may first be roughly aligned before aperture matching can take place.

4 Summary

The recovery of shape from a motion sequence requires in general establishing correspondence between the points in the images. This task is particularly difficult when the images are taken from viewpoints that are relatively distant from one another, conditions referred to as "long range motion". Establishing point correspondence under these conditions is important for constructing both object-centered as well as viewer-centered representations for object recognition. Such representations tend to be more stable as the images from which they are constructed are separated by relatively large transformations.

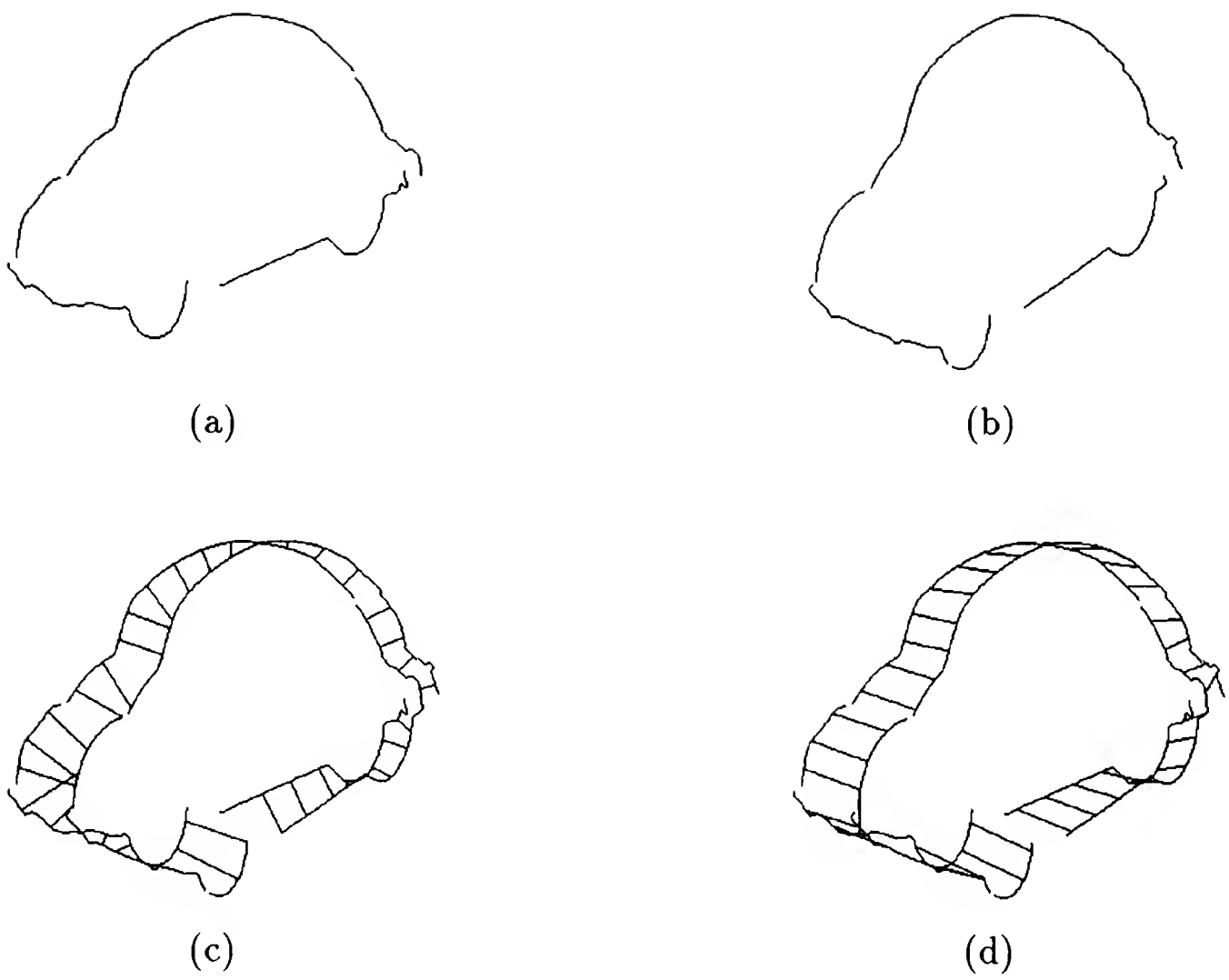


Figure 5: Matching silhouettes of a VW car under aperture conditions versus to using epipolar lines. (a,b) Two silhouette pictures of a VW car from two different viewpoints. (c) Matching the two silhouettes under aperture conditions. (d) Matching the two silhouettes by epipolar lines. (In these two figures the two silhouette drawings are overlapped. Straight lines connect matching points.)

Information about the shape of objects and the transformations they undergo can be used to guide the matching process. In this paper we reviewed the constraints imposed on the correspondence by rigid transformations and extended them to include images of objects that undergo general 3-D affine transformations as well as rigid objects with smooth surfaces. In all these cases the images are divided into epipolar lines, their correspondence is determined by the transformation, but the correspondence of points within the lines cannot be recovered. The epipolar lines can be computed from a small set of anchor points.

The correspondence is not determined uniquely even when three or more images are considered. Additional images can be used, however, in a heuristic algorithm to determine point correspondence. Such an algorithm is the trinocular stereovision algorithm [Yachida 1986], which is designed to work with sparse images and in the absence of uniform texture. We extended this algorithm to handle arbitrarily curved edge images and applied it to images of natural objects. We discussed the use of other constraints such as ordering, parallelism, and symmetry in solving the correspondence problem. Finally, we showed that epipolar lines can be used to improve matching obtained under aperture conditions. The techniques described in this paper were implemented and used to construct viewer-centered models for object recognition.

Acknowledgments I wish to thank Tao Alter,, Eric Grimson, Yael Moses, Tomaso Poggio, Amnon Shashua, and Shimon Ullman for their helpful comments throughout this work.

Appendix A

In this appendix we show how the transformation can be recovered (up to a scale factor) from two images under perspective projection using eight corresponding points. This is a repetition of the method presented in [Tsai and Huang 1984].

Let P_1 and P_2 be two perspective images of a rigid object obtained by a rotation R and a translation t in 3-D space. Denote r_x, r_y, and r_z the three row vectors of R, and (t_x, t_y, t_z) the three translation components. Note: we can substitute R in this analysis with any 3×3 matrix.

We define

$$\begin{aligned} \mathrm{a} &= t_y \mathrm{r}_x - t_x \mathrm{r}_y \\ \mathrm{b} &= t_y \mathrm{r}_z + t_z \mathrm{r}_y \\ \mathrm{c} &= t_z \mathrm{r}_x - t_x \mathrm{r}_z \end{aligned}$$

Note that a, b, and c are vectors in $\mathcal{R}^3$.

Let $(x_i, y_i) \in P_1$ and $(x'_i, y'_i) \in P_2$ be a pair of corresponding points, denote $\mathrm{p}_i = (x_i, y_i, 1)$, the following equations holds

$$\mathrm{a}\mathrm{p}_i = \mathrm{b}\mathrm{p}_i x'_i - \mathrm{c}\mathrm{p}_i y'_i$$

When anchor points are given, p_i, x'_i, and y'_i are known, while the vectors a, b, and c are not. These vectors contain nine components, and the equation is linear and homogeneous in their components. Therefore, a, b, and c can be recovered up to a scale factor using eight anchor points. Once the system is solved we can recover the parameters of the transformation (up to a scale factor in the translation components) using the following identities

$$\begin{aligned} \mathrm{a}^2 &= t_x^2 + t_y^2 \\ \mathrm{b}^2 &= t_y^2 + t_z^2 \\ \mathrm{c}^2 &= t_x^2 + t_z^2 \end{aligned}$$

And

$$\begin{aligned} \mathrm{ab} &= t_x t_z \\ \mathrm{ac} &= t_y t_z \\ \mathrm{bc} &= t_x t_y \end{aligned}$$

The translation components are therefore given (up to a scale factor) by

$$\begin{aligned} t_x^2 &= \frac{1}{2}(\mathrm{a}^2 - \mathrm{b}^2 + \mathrm{c}^2) \\ t_y^2 &= \frac{1}{2}(\mathrm{a}^2 + \mathrm{b}^2 - \mathrm{c}^2) \\ t_z^2 &= \frac{1}{2}(-\mathrm{a}^2 + \mathrm{b}^2 + \mathrm{c}^2) \end{aligned}$$

And the rotation matrix can be retrieved from

$$R = \begin{pmatrix} t_y & -t_x & 0 \\ 0 & t_z & t_y \\ t_z & 0 & -t_x \end{pmatrix}^{-1} \begin{pmatrix} \mathrm{a} \\ \mathrm{b} \\ \mathrm{c} \end{pmatrix}$$

Note that the rotation obtained is not scaled.

References

Ayache, N. and Lustman, F., 1987. Fast and reliable passive trinocular stereovision. *Proc. of the Int. Conf. on Computer Vision, London, UK*, pp. 422-427.

Baker, H.H. and Binford, T.O., 1981. Depth from edges and intensity based stereo. *Proceedings of the Int. Joint Conf. on Artificial Intelligence, Vancouver, BC*, pp. 631-636.

Bachelder, I.A. and Ullman, S., 1991. Contour matchingusing local affine transformation. M.I.T., A.I. Memo 1326, in press.

Basri, R. and Ullman, S., 1988. The alignment of objects with smooth surfaces. *Proc. of 2nd Int. Conf. on Computer Vision, Tampa, FL*, pp. 482-488.

Burt, P.J., Bergen, J., Hingorani, R., Peleg, S., and Anandan, P., 1990. Dynamic analysis of image motion for vehicle guidance. *Proc. of IEEE Int. Workshop on Intelligent Motion Control*, 75-82.

Bolles, R.C., Baker, H.H., and Marimont, D.H., 1987. Epipolar-plane image analysis: an approach to determining structure from motion. *The Int. J. of Computer Vision, Vol. 1.*

Fischler, M.A. and Bolles, R.C., 1981. Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography. *Communications of the ACM, Vol. 24, No. 6*, pp. 381-395.

Grimson, W.E.L., 1980. A computer implementation of a theory of human stereo vision. *Philosophical Trans. of the Royal Society, London. B 292*, pp. 217-253.

Grimson, W.E.L. and Lozano-Pérez, T., 1984. Model-based recognition and localization from sparse data. *Int. J. of Robotics Research, Vol. 3*, pp. 3-35.

Grimson, W.E.L. and Lozano-Pérez T., 1987. Recognition of object families using parameterized models, *Proc. of the First Int. Conf. on Computer Vision, IEEE Computer Society Press*, pp. 93-100.

Hildreth, E.C., 1984. Computations underlying the measurements of visual motion. *Artificial Intelligence, Vol. 23*, pp. 309-354.

Huang, T.S. and Lee, C.H., 1989. Motion and Structure from Orthographic Projections. *IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 2, No. 5*, pp. 536-540.

Huttenlocher, D.P. and Ullman, S., 1987. Object recognition using alignment. *Proc. of the Int. Conf. on Computer Vision, London, UK*, pp. 102-111.

Jacobs, D.W., 1991. *To be published.*

Koenderink, J.J. and Van Doorn, A.J., 1975. Invariant properties of the motion parallax field due to the movement of rigid bodies relative to an observer. *Optica Acta, Vol. 22, No. 9*, pp. 771-773.

Koenderink J.J. and Van Doorn A.J., 1979. The internal representation of solid shape with respect to vision. *Biological Cybernetics, Vol. 32*, pp. 211-216.

Lee, C.H. and Huang, T.S., 1990. Finding point correspondences and determining motion of a rigid object from two weak perspective views. *Computer Vision, Graphics, and Image Processing, Vol. 52*, pp. 309-327.

Longuet-Higgins, H.C., 1981. A computer algorithm for reconstructing a scene from two projections. *Nature, Vol. 293*, pp. 133-135.

Lowe, D.G., 1987. Three-dimensional object recognition from single two dimensional images. *Artificial intelligence Journal, Vol. 31*, pp. 355-395.

Marr D., 1977. Analysis of occluding contour. *Philosophical Trans. of the Royal Society, London, B 275*, pp. 483-524.

Marr, D. and Poggio, T., 1979. A computational theory of human stereo vision. *Proc. of the Royal Society, London. B 204*, pp. 301-328.

Marr, D. and Ullman, U., 1981. Directional selectivity and its use in early visual processing. *Proc. of the Royal Society, London. B 211*, pp. 151-180.

Poggio, T., 1990. 3D object recognition: on a result by Basri and Ullman. *Technical Report #9005-03, IRST*, Povo, Italy.

Poggio, T. and Edelman, S., 1990. A network that learns to recognize three dimensional objects. *Nature, Vol. 343*, pp. 263-266.

Roach, J.W. and Aggarwal, J.K., 1979. Computer tracking of objects moving in space. *IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 1, No. 2*, pp. 127-135.

Tsai, R.Y. and Huang, T.S., 1984. Uniqueness and estimation of three-dimensional motion parameters of rigid objects with curved surfaces. *IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 6, No. 1*, pp. 13-27.

Ullman, S., 1978. Two dimensionality of the correspondence process in apparent motion. *Perception 7*, pp. 683-693.

Ullman, S., 1979. The Interpretation of visual motion. *M.I.T. Press, Cambridge, MA.*

Ullman, S., 1983. Recent computational studies in the interpretation of structure from motion. In: A. Rosenfeld and J. Beck (Eds.), *Human and Machine Vision. Academic Press, New York, NY.*

Ullman, S., 1984. Maximizing rigidity: the incremental recovery of 3-D structure from rigid and rubbery motion. *Perception, Vol. 13*, pp. 255-274.

Ullman, S. and Basri, R., 1991. Recognition by linear combinations of models. *IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 13, No. 10.*

Yachida M., 1986. 3-D data acquisition by multiple views. *The third International Symposium on Robotic Research. M.I.T. Press, Cambridge, MA*, pp. 11-18.